GPT-4o System Card

Year: 2024 | Citations: 2523 | Authors: OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh

---

Abstract

GPT-4o is an autoregressive omni model that accepts as input any combination of text, audio, image, and video, and generates any combination of text, audio, and image outputs. It's trained end-to-end across text, vision, and audio, meaning all inputs and outputs are processed by the same neural network. GPT-4o can respond to audio inputs in as little as 232 milliseconds, with an average of 320 milliseconds, which is similar to human response time in conversation. It matches GPT-4 Turbo performance on text in English and code, with significant improvement on text in non-English languages, while also being much faster and 50\% cheaper in the API. GPT-4o is especially better at vision and audio understanding compared to existing models. In line with our commitment to building AI safely and consistent with our voluntary commitments to the White House, we are sharing the GPT-4o System Card, which includes our Preparedness Framework evaluations. In this System Card, we provide a detailed look at GPT-4o's capabilities, limitations, and safety evaluations across multiple categories, focusing on speech-to-speech while also evaluating text and image capabilities, and measures we've implemented to ensure the model is safe and aligned. We also include third-party assessments on dangerous capabilities, as well as discussion of potential societal impacts of GPT-4o's text and vision capabilities.